correlate well with the resolution of the X-ray data, as shown below for the volume-based *Z* scores.

21.2.2.2.3. *Deviations from standard atomic volumes as a quality measure for protein crystal structures*

The observations that protein X-ray structures are at least as tightly packed as small-molecule crystals (Richards, 1974; Harpaz *et al.*, 1994) and that the packing density inside proteins displays very limited variation (Richards, 1974; Finney, 1975) suggest that atomic volumes or measures of atomic packing can be added to the list of parameters for assessing the quality of protein structures.

Packing and related measures have been used to compare structures of proteins derived by both X-ray diffraction and NMR spectroscopy. Ratnaparkhi *et al.* (1998) analysed pairs of protein structures for which both crystal and NMR structures were available. They found that the packing values of the NMR models displayed a much larger scatter than those of the corresponding crystal structures, suggesting that this is probably due to the fact that accurate values of the packing density cannot, at present, be obtained from NMR data. Similar conclusions were reached using measures of residue–residue contact area (Abagyan & Totrov, 1997).

Here, we describe the approach of Pontius *et al.* (1996), in which deviations from standard atomic volumes are used to assess the quality of a protein model, both overall and in specific regions.

The volumes occupied by atoms and residues inside proteins can be readily computed using the Voronoi method (1908), first applied to proteins by Richards (1974) and Finney (1975). This method uses the atomic positions of the molecular model, and the volume assigned to each atom is defined as the smallest polyhedron created by the set of planes bisecting the lines joining the atom centre to those of its neighbours (Fig. 21.2.2.1).

The use of the classical Voronoi procedure is justified in the context of validation because it avoids the need to derive a consistent set of van der Waals radii for atoms in the system. Such sets are used by other volume-calculation methods in order to partition space more accurately (Richards, 1974, 1985; Gellatly & Finney, 1982). Assigning a consistent set of radii to protein atoms is, indeed, not straightforward due to the heterogeneity of the interactions within the protein (polar, ionic, non-polar) and the presence of a large variety of hetero groups.

Structure-quality assessment based on volume calculations involves computing the atomic volumes in a subset of highly resolved and refined protein structures and analysing the distributions of these volumes for different atomic types, defined according to their chemical nature and bonded environment. These distributions define the expected ranges (mean and standard deviation) for the volume of each category of atoms. Atomic volumes in a given structure are then compared to the expected ranges, and statistically significant deviations from these ranges are flagged.

The program *PROVE* (Pontius *et al.*, 1996) implements such an approach using the analytic algorithms for volume and surface-area calculations encoded in *SurVol* (Alard, 1991). It computes for each atom i in a structure its volume Z score (Z score $= \left|V_i^k - \overline{V^k}\right|/\sigma^k$), where the superscript k designates the particular atom type (*e.g.*, the C^α atom in a Leu residue), and $\overline{V^k}$ and σ^k are, respectively, the mean and standard deviation of the reference volume distribution for the corresponding atom type. These reference distributions are derived from a set of high-quality protein crystal structures using exactly the same calculation procedure (Pontius *et al.*, 1996).

Atoms with absolute Z scores > 3 are flagged as possible problem regions in the protein model, and residues containing such atoms are highlighted on graphical plots of the same type as those used by the *PROCHECK* program and on molecular models displayed using programs such as *Rasmol* (Sayle & Milner-White, 1995).

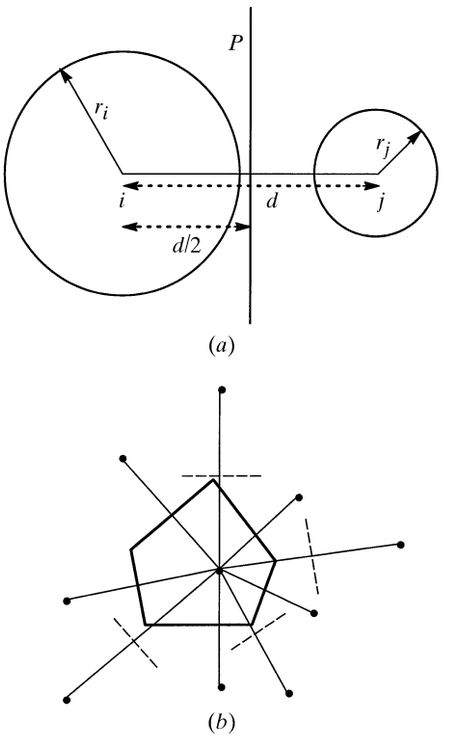


Fig. 21.2.2.1. The Voronoi polyhedron. (*a*) Positioning of the dividing plane P between two atoms i and j, with van der Waals radii r_i and r_j, respectively, separated by a distance d. The plane P is positioned at $d/2$. (*b*) 2D representation of the Voronoi polyhedron of the central atom. This polyhedron is the smallest polyhedron delimited by all the dividing planes of the atom.

In addition to the validation of the local quality of the model, its overall quality can be assessed by the *root-mean-square volume Z score* of all its atoms (see Fig. 21.2.2.2 for definition). As for many stereochemical global quality indicators, this Z score shows good correlation with the nominal resolution (d spacing) of the crystallographic data, as illustrated in Fig. 21.2.2.2(*a*). This figure also shows that Z-score ranges can be defined for each resolution interval. The Z scores of individual proteins that lie outside these intervals may be indicative of 'problem' structures. This is clearly the case for the two proteins 2ABX and 2GN5, whose Z scores are much higher than average (Fig. 21.2.2.2*b*).

Since the Voronoi volume of solvent-accessible atoms cannot be defined, because these atoms are not completely surrounded by other atoms, only completely buried atoms are scored.

The current version of *PROVE* is unable to measure the deviations from standard volumes for atoms in nucleic acids or hetero groups, simply because of the lack of reference volumes for these structures. This should change in the near future, at least for nucleic acids, thanks to the growing number of high-quality nucleic acid crystal structures from which standard volume ranges could be readily derived.

21.2.3. Validation of a model *versus* experimental data

By far the most important measure of the quality of a given atomic model is its agreement with the experimental data. This type of validation is geared towards detecting systematic errors, which determine the overall accuracy of the model, and random errors, which affect the precision of the model.

Systematic errors are difficult to detect even in highly refined structures, especially at lower resolution. The most commonly used

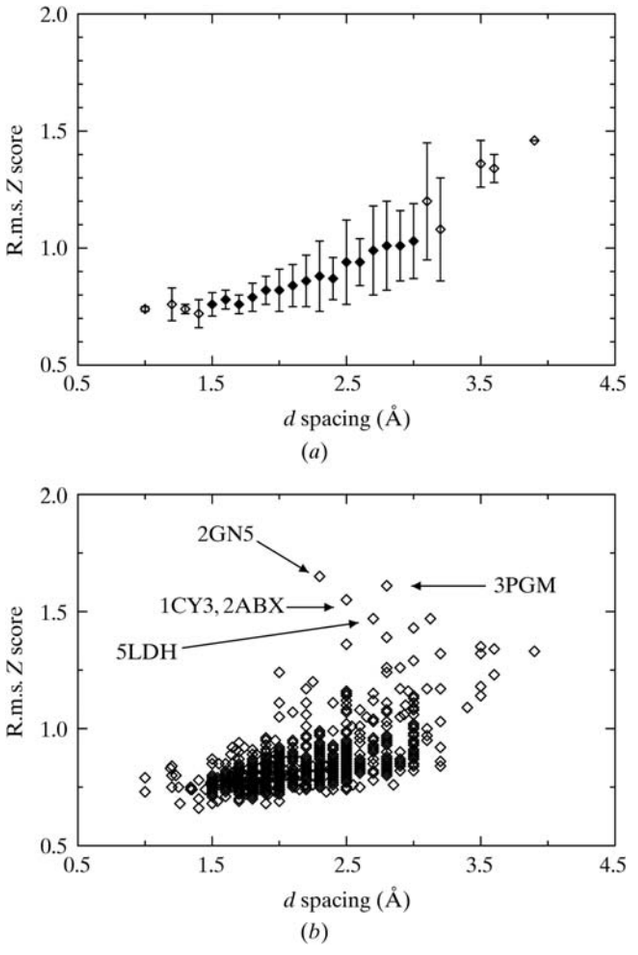


Fig. 21.2.2.2. Atomic volume Z score r.m.s. variation with nominal resolution (d spacing) in 900 protein structures from the PDB. (*a*) Average of the r.m.s. volume Z score computed for structures having the same resolution (to within ± 0.1 Å). The vertical bars indicate the magnitude of the standard deviations of the r.m.s. volume Z score in individual d-spacing bins. Graph points are derived from less than 10 structures (open diamonds) and from more than 10 structures (filled diamonds). (*b*) R.m.s. Z-score values as in (*a*), displayed for individual structures as a function of resolution. The five furthest outlier proteins are marked by their PDB codes.

measures of the agreement between the atomic coordinates and the X-ray data are the classical R factor and the 'free R factor' ($R_{\rm free}$) (Brünger, 1992*b*). The latter is based on standard statistical cross-validation techniques (Brünger, 1997) and is therefore less amenable to manipulation, such as leaving out weak data or over-fitting the data with too many parameters. Currently, nearly half of the publications on macromolecular structures report $R_{\rm free}$ values, an indication that its use is becoming more widespread. So far, however, there are no clear guidelines indicating what an 'acceptable' $R_{\rm free}$ value should be (Kleywegt & Brünger, 1996).

An expression for estimating the expected $R_{\rm free}$ value has been proposed (see Dodson *et al.*, 1996) and used to assess the significance of the drop in $R_{\rm free}$ during refinement. Accurate expressions for the expected ratio of $R_{\rm free}$ to R (the $R_{\rm free}$ ratio) have also been derived theoretically (Tickle *et al.*, 1998). This ratio seems to be independent of random errors and can be used to detect systematic errors at the convergence of the least-squares refinement. The remaining problem is to determine what the precision of $R_{\rm free}$ or the $R_{\rm free}$ ratio should be. In other words, if the $R_{\rm free}$ ratio differs from the expected value, when is the difference significant? This requires knowing the variance of these parameters. Estimating the precision of $R_{\rm free}$ can be done empirically by performing repeated refinements of the same structure with different sets of reflections removed (Brünger, 1997). From such analysis, a useful approximation to the $R_{\rm free}$ precision was suggested to be the ratio $R_{\rm free}/(n)^{1/2}$, where n is the number of reflections in the test set.

Evaluating the precision of the refined parameters, that is, the atomic coordinates and the temperature or B factors, is a different matter. In small-molecule crystallography, the standard uncertainty (s.u.) of the parameters can be computed from the variance–covariance matrix, obtained by inverting the full normal-equations matrix (Cruickshank, 1965). This can, in principle, also be done for the parameters of macromolecules. However, the number of second derivatives to be computed and the size of the matrix to be inverted are so large that this task is too time consuming to be performed routinely. This is gradually changing, however. An increasing number of proteins structures, primarily those solved at atomic resolution, have their s.u.'s computed in this manner (Deacon *et al.*, 1997; Harata *et al.*, 1998). A program often used for this purpose is *SHELXL* (Sheldrick & Schneider, 1997), a well known refinement software package for small molecules that has recently been extended to proteins. Availability of s.u.'s can determine the dependence of the precision of the atomic coordinates on various factors, such as the resolution, the atomic number, and the number and types of restraints used during refinement (Tickle *et al.*, 1998).

Other methods for determining the relative precision of atoms in macromolecular structures involve calculating the agreement between the model and the electron density in specific regions. The newer approach by Zhou *et al.* (1998) is related to the real-space R factor of Jones *et al.* (1991), but differs from it by the way in which the electron density is computed (Chapman, 1995).

As our understanding of the factors that govern the systematic errors in macromolecular crystallography increases and our ability to detect random errors improves, the possibility of devising systematic and possibly more automatic protocols for assessing the agreement between the model and the data will emerge.

In what follows, we describe the software package *SFCHECK* (Vaguine *et al.*, 1999), which can be regarded as a first attempt in this direction. This software computes and summarizes many of the commonly used measures for evaluating the quality of the structure-factor data and the agreement of the model with these data.

We summarize the tasks performed and the quality indicators computed by *SFCHECK* and briefly illustrate how this software can be used to evaluate individual structures and survey different structures.

21.2.3.1. *A systematic approach using the SFCHECK software*

21.2.3.1.1. *Tasks performed by SFCHECK*

21.2.3.1.1.1. *Treatment of structure-factor data and scaling*

SFCHECK reads in the structure-factor data written in mmCIF format. It then performs the following operations: Reflections are excluded if they are systematically absent, negative, or have flagged σ values (99.9). Equivalent reflections are merged. The amplitudes of missing reflections are approximated by taking the average value for the corresponding resolution shell.

From the model coordinates read from the PDB (or mmCIF) atomic coordinates file, *SFCHECK* calculates structure factors and scales them to the observed structure factors. The scaling factor, S, is computed using a smooth cutoff for low-resolution data (Vaguine *et al.*, 1999) (Table 21.2.3.1). This involves the calculation of the observed and calculated overall B factors from the standard deviations of the Gaussian fitted to the Patterson origin peaks [see Table 21.2.3.1 and Vaguine *et al.* (1999)]. In addition, *SFCHECK* also estimates the overall anisotropy of the data, following the

Table 21.2.3.1. *Parameters computed for the analysis of the structure-factor data*

The first column lists the parameter, the second column gives the formula or definition of the parameter and the third column contains a short description of the meaning of the parameters when warranted.

Parameter	Formula/definition	Meaning
Completeness (%)	Percentage of the expected number of reflections for the given crystal space group and resolution	
B_overall (Patterson)	$8\pi^2\sigma_{\rm Patt}/(2)^{1/2}$ *	Overall B factor
R_stand(F)	$\langle\sigma(F)\rangle/\langle F\rangle$ †	Uncertainty of the structure-factor amplitudes
Optical resolution	$(\sigma^2_{\rm Patt}+\sigma^2_{\rm sph})^{1/2}$ *‡	Expected minimum distance between two resolved atomic peaks
Expected optical resolution	Optical resolution computed considering all reflections	
CC_F	$\dfrac{\langle F_{\rm obs}F_{\rm calc}\rangle-\langle F_{\rm obs}\rangle\langle F_{\rm calc}\rangle}{\left[(\langle F^2_{\rm obs}\rangle-\langle F_{\rm obs}\rangle^2)(\langle F^2_{\rm calc}\rangle-\langle F_{\rm calc}\rangle^2)\right]^{1/2}}$	Correlation coefficient between the observed and calculated structure-factor amplitudes
S	$\left\{\dfrac{\sum(F_{\rm obs}f_{\rm cutoff})^2}{\sum\left[F_{\rm calc}\exp(-B^{\rm overall}_{\rm diff}s^2)f_{\rm cutoff}\right]^2}\right\}^{1/2}$ §	Factor applied to scale $F_{\rm calc}$ to $F_{\rm obs}$
$f_{\rm cutoff}$	$1-\exp(-B_{\rm off}s^2)$ ¶	Function applied to obtain a smooth cutoff for low-resolution data

* $\sigma_{\rm Patt}$ is the standard deviation of the Gaussian fitted to the Patterson origin peak.
† F is the structure-factor amplitude, and $\sigma(F)$ is the structure-factor standard deviation. The brackets denote averages.
‡ $\sigma_{\rm sph}$ is the standard deviation of the spherical interference function, which is the Fourier transform of a sphere of radius $1/d_{\rm min}$, with $d_{\rm min}$ being the minimum d spacing.
§ $B^{\rm overall}_{\rm diff}=B^{\rm overall}_{\rm obs}-B^{\rm overall}_{\rm calc}$ is added to the calculated overall B factor, $B_{\rm overall}$, so as to make the width of the calculated Patterson origin peak equal to the observed one; s is the magnitude of reciprocal-lattice vector.
¶ $B_{\rm off}=4d^2_{\rm max}$, where s and $d_{\rm max}$, respectively, are the magnitude of the reciprocal-lattice vector and the maximum d spacing.

approach of Sheriff & Hendrickson (1987), and applies the anisotropic scaling after the Patterson scaling is performed (Murshudov *et al.*, 1998).

To assess the quality of the structure-factor data, the program computes four additional quantities (see Table 21.2.3.1 for details): the completeness of the data, the uncertainty of the structure-factor amplitudes, the optical resolution and the expected optical resolution. The latter two quantities represent the expected minimum distance between two resolved atomic peaks in the electron-density map when the latter is computed with the set of reflections specified by the authors and with all the reflections, respectively.

21.2.3.1.1.2. *Global agreement between the model and experimental data*

To evaluate the global agreement between the atomic model and the experimental data, the program computes three classical quality indicators: the R factor, $R_{\rm free}$ (Brünger, 1992*b*) and the correlation coefficient CC_F between the calculated and observed structure-factor amplitudes (Table 21.2.3.1). The R factor is computed using all the reflections considered (except those approximated by their average value in the corresponding resolution shell) and applying the same resolution and σ cutoff as those reported by the authors. $R_{\rm free}$ is computed using the subset of reflections specified by the authors. In addition, the R factor is evaluated using the 'non-free' subset of reflections (those not used to compute $R_{\rm free}$). The correlation coefficient is computed using all reflections from the reported high-resolution limit, applying the smooth low-resolution cutoff (see Table 21.2.3.1) but no σ cutoff.

21.2.3.1.1.3. *Estimations of errors in atomic positions*

The errors associated with the atomic positions are expressed as standard deviations (σ) of these positions. *SFCHECK* computes three different error measures. One is the original error measure of Cruickshank (1949). The second is a modified version of this error measure, in which the difference between the observed and calculated structure factors is replaced by the error in the experimental structure factors. The first two error measures are the expected maximal and minimal errors, respectively, and the third measure is the diffraction-component precision indicator (DPI). The mathematical expressions for these error measures are given in Table 21.2.3.2, and further details can be found in Vaguine *et al.* (1999).

21.2.3.1.1.4. *Local agreement between the model and the experimental data*

In addition to the global structure quality measures, *SFCHECK* also determines the quality of the model in specific regions. Several quality estimators can be calculated for each residue in the macromolecule and, whenever appropriate, for solvent molecules and groups of atoms in ligand molecules. These estimators are the normalized atomic displacement (Shift), the correlation coefficient between the calculated and observed electron densities (Density correlation), the local electron-density level (Density index), the average B factor (B-factor) and the connectivity index (Connect), which measures the local electron-density level along the molecular backbone. These quantities are computed for individual atoms and averaged over those composing each residue or group of atoms [see Table 21.2.3.3 and Vaguine *et al.* (1999) for details].

21.2.3.1.2. *Evaluation of individual structures*

Figs. 21.2.3.1–21.2.3.3 summarize the analysis carried out by *SFCHECK* on the protein rusticyanin from *Thiobacillus ferrooxidans* (1RCY) (Walter *et al.*, 1996). Fig. 21.2.3.1 displays the numerical results from the analysis of the structure-factor data and from the evaluation of the global agreement between the model and the data. The R-factor and $R_{\rm free}$ values, computed by *SFCHECK*

Table 21.2.3.2. *Estimation of errors in atomic coordinates*

The first column lists the parameter, the second column gives the formula or definition of the parameter and the third column contains a short description of the meaning of the parameters when warranted.

Parameter	Formula/definition	Meaning
$\sigma(x)$	$\dfrac{\sigma(\text{slope})}{\text{curvature}}$ *	Standard deviation of the atomic coordinates following Cruickshank (1949) for the minimal and maximal errors (Vaguine *et al.*, 1999)
σ(slope) for maximal error	$\dfrac{2\pi\left\{\sum\left[h^2(F_{\text{obs}}-F_{\text{calc}})^2\right]\right\}^{1/2}}{V_{\text{unit cell}}a}$ †	Expression for σ(slope) in the expected maximal error following Cruickshank (1949)
Curvature	$\dfrac{2\pi\sum(h^2F_{\text{obs}})}{V_{\text{unit cell}}a^2}$	Expression for the curvature following Murshudov *et al.* (1997)
σ(slope) for minimal error	$\dfrac{2\pi^2\left\{\sum\left[h^2\sigma(F_{\text{obs}})^2\right]\right\}^{1/2}}{V_{\text{unit cell}}a}$ ‡	Expression for σ(slope) in the expected minimal error, following Cruickshank (1949)
DPI	$\sigma(x)=\left(\dfrac{N_{\text{atoms}}}{N_{\text{obs}}-4N_{\text{atoms}}}\right)^{1/2}c^{-1/3}d_{\min}R$ §	Atomic coordinate error estimate following Cruickshank (1996)

* σ(slope) and curvature are the slope and curvature of the electron-density map at the atomic centre, in the x direction, for spherically symmetric peaks; $\sigma(x)\simeq\sigma(y)\simeq\sigma(z)$.
† a is the crystal unit-cell length, h is the Miller index and $V_{\text{unit cell}}$ the unit-cell volume.
‡ $\sigma(F_{\text{obs}})$ is the standard deviation of the structure-factor amplitude.
§ c is the structure-factor data completeness expressed as a fraction (0–1), R is the conventional R factor, N_{atoms} is the total number of atoms in the unit cell, N_{obs} is the total number of observed reflections and $d_{\min}$ is the minimum d spacing.

(Model vs. Structure Factors panel) using the identical reflection subset to that reported by the authors (Refinement panel), show negligible differences with the reported values. These differences are 0.175 *versus* 0.172 for the R factor and 0.25 *versus* 0.243 for R_{free}. The small R-factor difference may stem from the fact that *SFCHECK* considers a somewhat different number of reflections (9144) than the authors (9098), although it uses the same d-spacing range and σ cutoff as those reported.

The information in Figs. 21.2.3.1 and 21.2.3.2 allows one to make some judgement about the quality of the structure-factor data for this protein. The relatively high resolution of this structure (1.9 Å) is accompanied by limited data completeness (82.1%). The Rstand(F) plot on the same graph shows, furthermore, a decrease in quality of the high-resolution data (2.2–1.9 Å). The average radial completeness plot (bottom left-hand plot of Fig. 21.2.3.2) allows one to identify the regions in reciprocal space with incomplete data.

Fig. 21.2.3.3 presents the *SFCHECK* analysis of the local agreement of the model with the electron density for 1RCY. The shift plot shows that both backbone and side-chain shifts are of comparable size, with several residues (1, 2, 16, 25) displaying shifts as high as 0.16 Å. The density correlation is excellent throughout the entire molecule, except for residues 2, 16 and 29,

Table 21.2.3.3. *Parameters computed by SFCHECK to assess the quality of the model in specific regions*

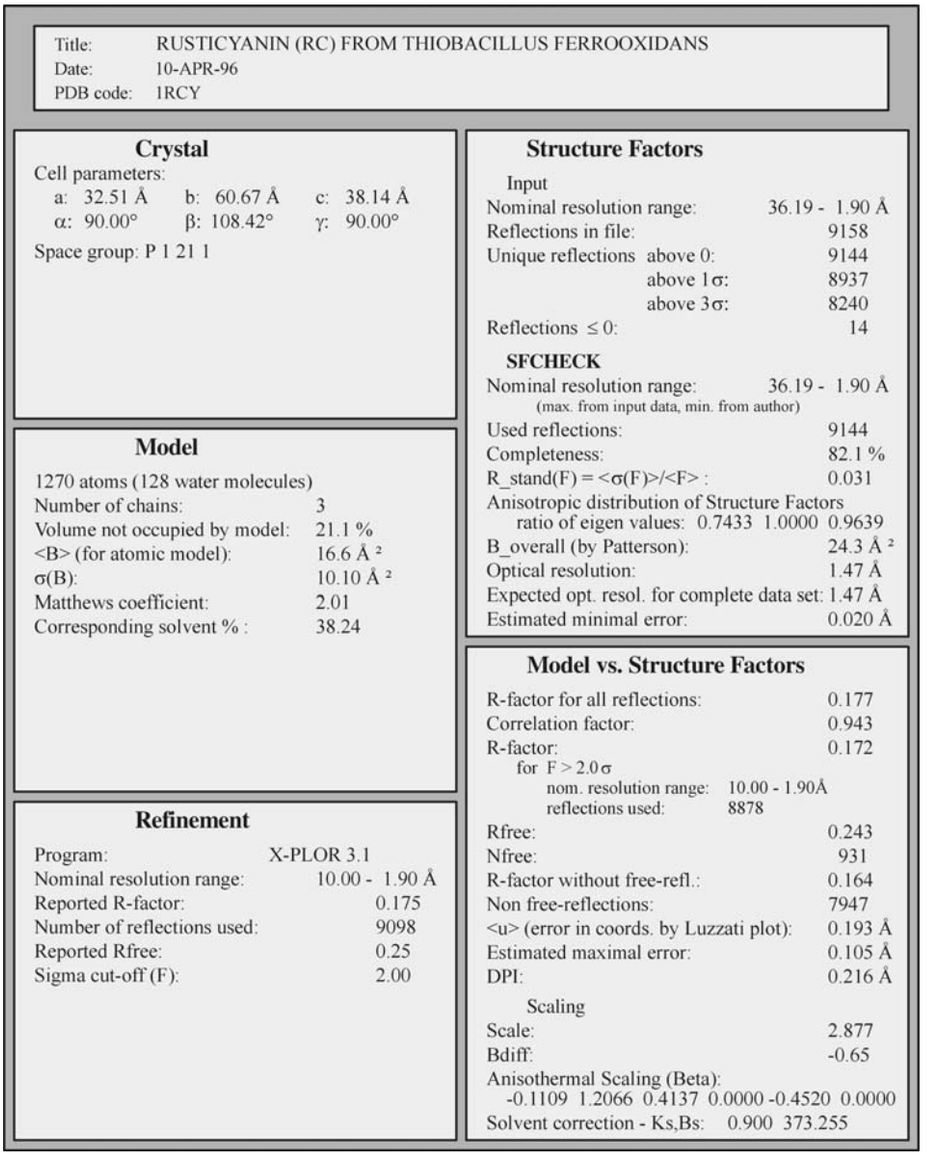
The first column lists the parameter, the second column gives the formula or definition of the parameter and the third column contains a short description of the meaning of the parameters when warranted.

Parameter	Formula/definition	Meaning
Shift	$(1/N\sigma)\sum_i^N\Delta_i,\ \text{with}\ \Delta_i=(\text{gradient}_i/\text{curvature}_i)$ *	Normalized average atomic displacement computed over a group of atoms or residue; reflects the tendency of the group of atoms to move from their current position
Density correlation	$\dfrac{\sum\rho_{\text{calc}}(x_i)[2\rho_{\text{obs}}(x_i)-\rho_{\text{calc}}(x_i)]}{\left(\left[\sum\rho^2_{\text{calc}}(x_i)\right]\left\{\sum[2\rho_{\text{obs}}(x_i)-\rho_{\text{calc}}(x_i)]^2\right\}\right)^{1/2}}$ †	Electron density correlation coefficient computed over a group of atoms or residue; reflects the local agreement of the model with the electron density
Density index	$[\prod\rho(x_i)]^{1/N}/\langle\rho\rangle_{\text{all atoms}}$ ‡	Reflects the level of the electron density for a group of atoms; is a local measure of the density level
Connect		Same as Density index, but considering only backbone atoms.§

* Gradient$_i$ is the gradient of the $F_{\text{obs}}-F_{\text{calc}}$ map with respect to the atomic coordinates, curvature$_i$ is the curvature of the model map computed at the atomic centre (see Agarwal, 1978), N is the number of atoms in the group considered and σ is the standard deviation of the Δ_i values computed in the structure.
† $\rho_{\text{calc}}(x_i)$ and $\rho_{\text{obs}}(x_i)$ are, respectively, the electron density computed from calculated and observed structure-factor amplitudes at the atomic centre. The summation is performed over all the atoms in the group considered. For polymer residues, D_corr is computed separately for backbone and side-chain atoms. For the calculation of the electron density at the atomic centre, see Vaguine *et al.* (1999).
‡ $[\prod\rho(x_i)]^{1/N}$ is the geometric mean of the $2F_{\text{obs}}-F_{\text{calc}}$ electron density of the atom subset considered and $\langle\rho\rangle_{\text{all atoms}}$ is the average electron density of the atoms in the structure. For water molecules or ions which are represented by a unique atom, the above expression reduces to the ratio $\rho(x_i)/\langle\rho\rangle_{\text{all atoms}}$.
§ Backbone atoms are N, C, C^{α}, for proteins and P, O5′ C5′ C3′ O3′ for nucleic acids.

which display poorer correlation. In particular, the side chains of these residues seem to be more poorly defined in the electron-density map. The backbone density index plunges in a few regions, notably at the N-terminus (residues 5–7) and in the segments comprising residues 25–30 and 68–70. The side chains display, in general, a poorer density index than the backbone, with some regions (for example, residues 5–7, 23–30, 58–60) displaying rather low density indices. The same segments also display higher backbone and side-chain *B* factors. The backbone Connect parameter is, on the other hand, quite good throughout, except for residues 5–7 and 28–29 (Fig. 21.2.3.3).

Water molecules (labeled w in the *SFCHECK* output) are also evaluated. The relevant plots for these molecules are those of the Shift, Density index and B factor parameters. The first 50 or so water molecules in the list (appearing sequentially along the plot from left to right) tend to display a higher density index and lower *B* factors (< 30 Å^2) than the following molecules in the list. They thus seem to be more reliably positioned than subsequent molecules, whose density indices sometimes drop perilously. A steady climb of the *B* factors is also apparent as one goes down the list of water molecules. The analysis of the density indices and *B* factors of individual water molecules performed by *SFCHECK* could be a very useful guide in investigations of the properties of crystallographic water molecules and their interactions with protein atoms.



Title: RUSTICYANIN (RC) FROM THIOBACILLUS FERROOXIDANS
Date: 10-APR-96
PDB code: 1RCY

Crystal

Cell parameters:
a: 32.51 Å b: 60.67 Å c: 38.14 Å
α: 90.00° β: 108.42° γ: 90.00°
Space group: P 1 21 1

Model

1270 atoms (128 water molecules)

Number of chains:	3
Volume not occupied by model:	21.1 %
<B> (for atomic model):	16.6 Å ²
σ(B):	10.10 Å ²
Matthews coefficient:	2.01
Corresponding solvent % :	38.24

Refinement

Program:	X-PLOR 3.1
Nominal resolution range:	10.00 - 1.90 Å
Reported R-factor:	0.175
Number of reflections used:	9098
Reported Rfree:	0.25
Sigma cut-off (F):	2.00

Structure Factors

Input

Nominal resolution range:	36.19 - 1.90 Å
Reflections in file:	9158
Unique reflections above 0:	9144
above 1σ:	8937
above 3σ:	8240
Reflections ≤ 0:	14

SFCHECK

Nominal resolution range: (max. from input data, min. from author)	36.19 - 1.90 Å
Used reflections:	9144
Completeness:	82.1 %
R_stand(F) = <σ(F)>/<F> :	0.031
Anisotropic distribution of Structure Factors ratio of eigen values: 0.7433 1.0000 0.9639	
B_overall (by Patterson):	24.3 Å ²
Optical resolution:	1.47 Å
Expected opt. resol. for complete data set:	1.47 Å
Estimated minimal error:	0.020 Å

Model vs. Structure Factors

R-factor for all reflections:	0.177
Correlation factor:	0.943
R-factor:	0.172
for F > 2.0 σ	
nom. resolution range:	10.00 - 1.90Å
reflections used:	8878
Rfree:	0.243
Nfree:	931
R-factor without free-refl.:	0.164
Non free-reflections:	7947
<u> (error in coords. by Luzzati plot):	0.193 Å
Estimated maximal error:	0.105 Å
DPI:	0.216 Å

Scaling

Scale:	2.877
Bdiff:	-0.65

Anisothermal Scaling (Beta):
-0.1109 1.2066 0.4137 0.0000 -0.4520 0.0000
Solvent correction - Ks,Bs: 0.900 373.255

Fig. 21.2.3.1. Typical *SFCHECK* output in PostScript format, illustrated for the protein rusticyanin from *Thiobacillus ferrooxidans* (1RCY) (Walter *et al.*, 1996). Summary panels displaying the numerical results from the analysis of the deposited structure-factor data and from the evaluation of the global agreement between the model and these data. The top elongated panel lists the PDB title record, deposition date and PDB code. The Crystal panel summarizes the crystal parameters, provided by the authors, as read from the model input files. The Model and Refinement panels list the information provided by the authors on the model and the refinement procedure, respectively. This information is read from the PDB coordinates entry. The Structure Factors panel summarizes the information on the deposited structure-factor data (Input section) and on the data used and criteria computed by *SFCHECK* (SFCHECK section). The numbers given under 'Anisotropic distribution of Structure Factors' are the ratios of the eigenvalues of the symmetric anisotropic thermal tensor to the maximum eigenvalues. The Model vs. Structure Factors panel summarizes the results of the verifications made by *SFCHECK*. The values listed under 'Anisothermal Scaling (Beta)' are those of the overall anisotropic thermal tensor ($b_{11}, b_{12}, b_{13}, b_{22}, b_{23}, b_{33}$). The meanings of other listed quantities are either self-explanatory or are described in the text.

21.2.3.1.3. *Quality assessment based on surveys across structures*

21.2.3.1.3.1. *Assessing the quality of a structure as a whole*

As for the evaluation of the geometric and stereochemical parameters of the model, surveying the same quality indicators across many structures is crucial. It allows one to establish the ranges of expected values for each indicator and to identify structures with unexpected features – those for which the values of one or more quality indicators are outside their standard range.

The global quality indicators computed by *SFCHECK* are the nominal resolution (*d* spacing), the *R* factor, $R_{\rm free}$, the minimal and maximal errors in atomic positions, the DPI, and the correlation coefficient CC_F. Another type of global quality indicator can be obtained by computing the average values of local quality measures across a given structure. This can be done for the per-residue (or per-group) atomic displacement and the Density correlation and B factor parameters as well as for the Density index and Connect parameters.

Many of the geometric and stererochemical quality indicators vary as a function of resolution – some linearly and some not (Laskowski *et al.*, 1993). This is also the case for most of the global quality indicators described here. Examples of this dependence are given in Fig. 21.2.3.4, which shows how the correlation coefficient, the maximal error, the average atomic displacement and average density index vary as a function of resolution in the 104 nucleic acid structures surveyed. This variation is approximately linear for all four parameters. The density correlation and average density index decrease, whereas the maximal error and average atomic displacements increase, as the resolution gets poorer. In all four plots of

Fig. 21.2.3.4, the points tend to display significant scatter as the *d* spacing increases, and at least three points, corresponding to the same three structures, appear as outliers in all plots. These structures also appear as outliers in the analysis of other parameters. A closer examination revealed that in the vast majority of the cases, the abnormal behaviour of these structures could be traced back to problems with data formats or errors that occurred during data deposition and entry processing.

As the number of structures with deposited structure-factor data becomes large enough, plots such as those of Fig. 21.2.3.4 could be used to define the expected range of values for a quality indicator in a structure determined at a given resolution or refined under given conditions. Structures yielding quality indicators outside this range could then be identified as unusual on a more solid statistical basis.

21.2.3.1.3.2. *Assessing the quality in specific regions of a model*

The main purpose for computing the four local quality measures, the B factor, the Density index, the atomic displacement (Shift) and the Density correlation (Table 21.2.3.3), is to identify problem regions in a model. In order to do this effectively, it is necessary to evaluate the degree of redundancy between these measures and to establish the standard ranges for their values. The latter task, in particular, is not straightforward since it depends crucially on the quality of the experimental data and biases introduced by the scaling procedure and refinement protocol. In this regard, several issues are presently still under investigation.

A preliminary investigation of the mutual relations between the above-mentioned local measures has been performed in several protein and nucleic acid structures taken individually. This shows that that the *B* factor is strongly correlated with the density index, as illustrated in Fig. 21.2.3.5(*a*), and to a lesser extent with the atomic displacement (Fig. 21.2.3.5*b*). A weaker correlation was detected between the latter three measures and the residue density correlation (data not shown).

Analyses across structures could, in principle, be carried out for all four local measures computed by *SFCHECK*, provided these measures are not subject to systematic biases due to differences in

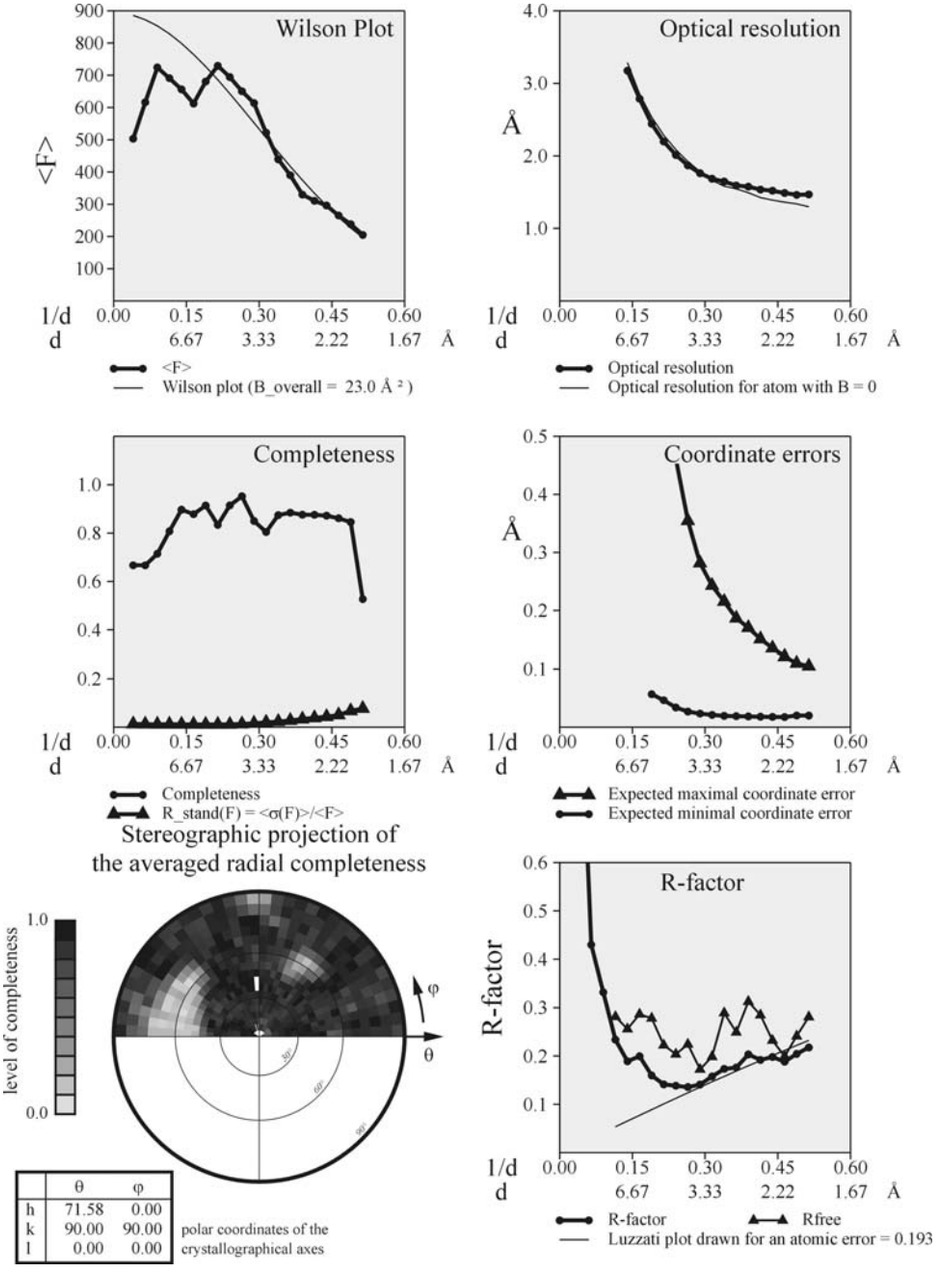


Fig. 21.2.3.2. Graphical output from the *SFCHECK* analysis of global characteristics of the structure-factor data and the model agreement with those data for the same structure as in Fig. 21.2.3.1. From left to right and top to bottom: the Wilson plot; the behaviour of the optical resolution as a function of the nominal resolution (*d* spacing); the data completeness and structure-factor standard error as a function of the *d* spacing; the maximal and minimal coordinate error dependence on *d* spacing; a stereographic projection of the averaged radial structure-factor data completeness; and, finally, the *R*-factor dependence and Luzzati plots for a given atomic error.

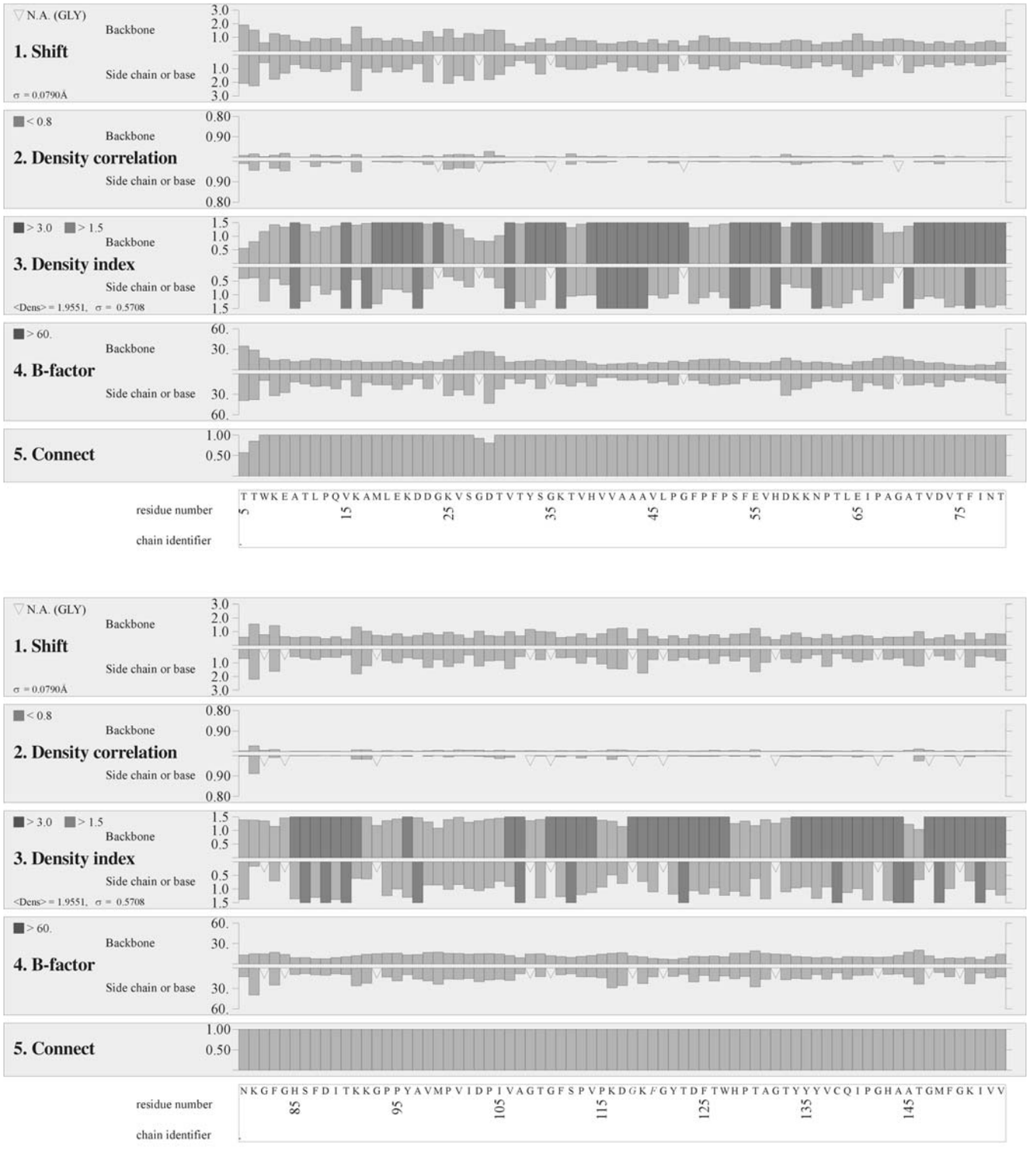


Fig. 21.2.3.3. *SFCHECK* evaluation summary of the local agreement between the model and the electron density for the same structure as in Fig. 21.2.3.1. Five criteria are plotted for each residue of the macromolecule (designated by its one-letter code), as well as for each solvent molecule (w), or hetero group. These criteria are: (1) Shift, (2) Density correlation, (3) Density index, (4) B factor, (5) Connect. The definitions of these criteria are given in the text. Note that the values of the Connect parameter are truncated to a maximum of 1. The *SFCHECK* output shown in Figs. 21.2.3.1–21.2.3.3 was generated using routines from *PROCHECK* kindly provided by R. Laskowski.

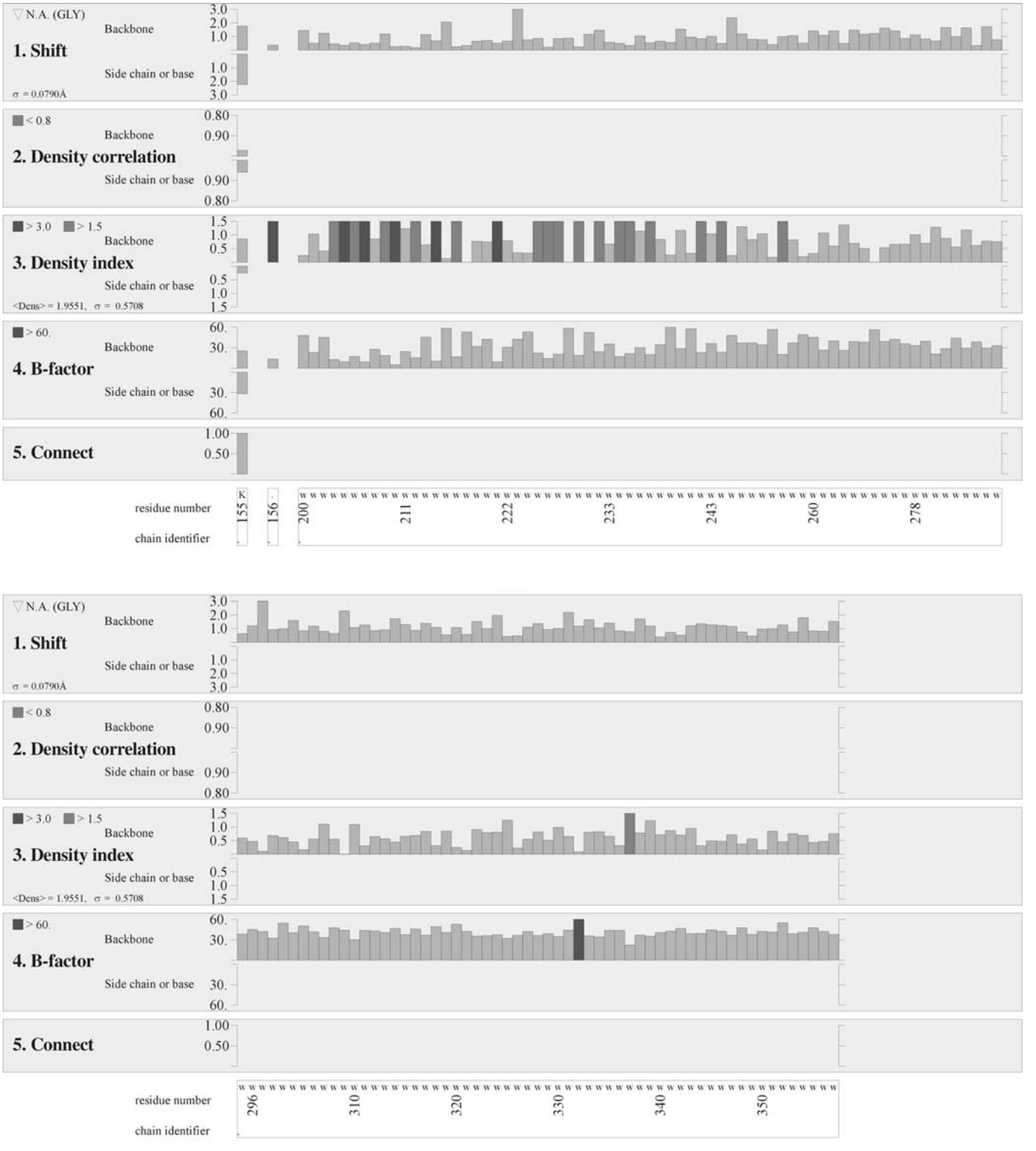


Fig. 21.2.3.3. (*cont.*)

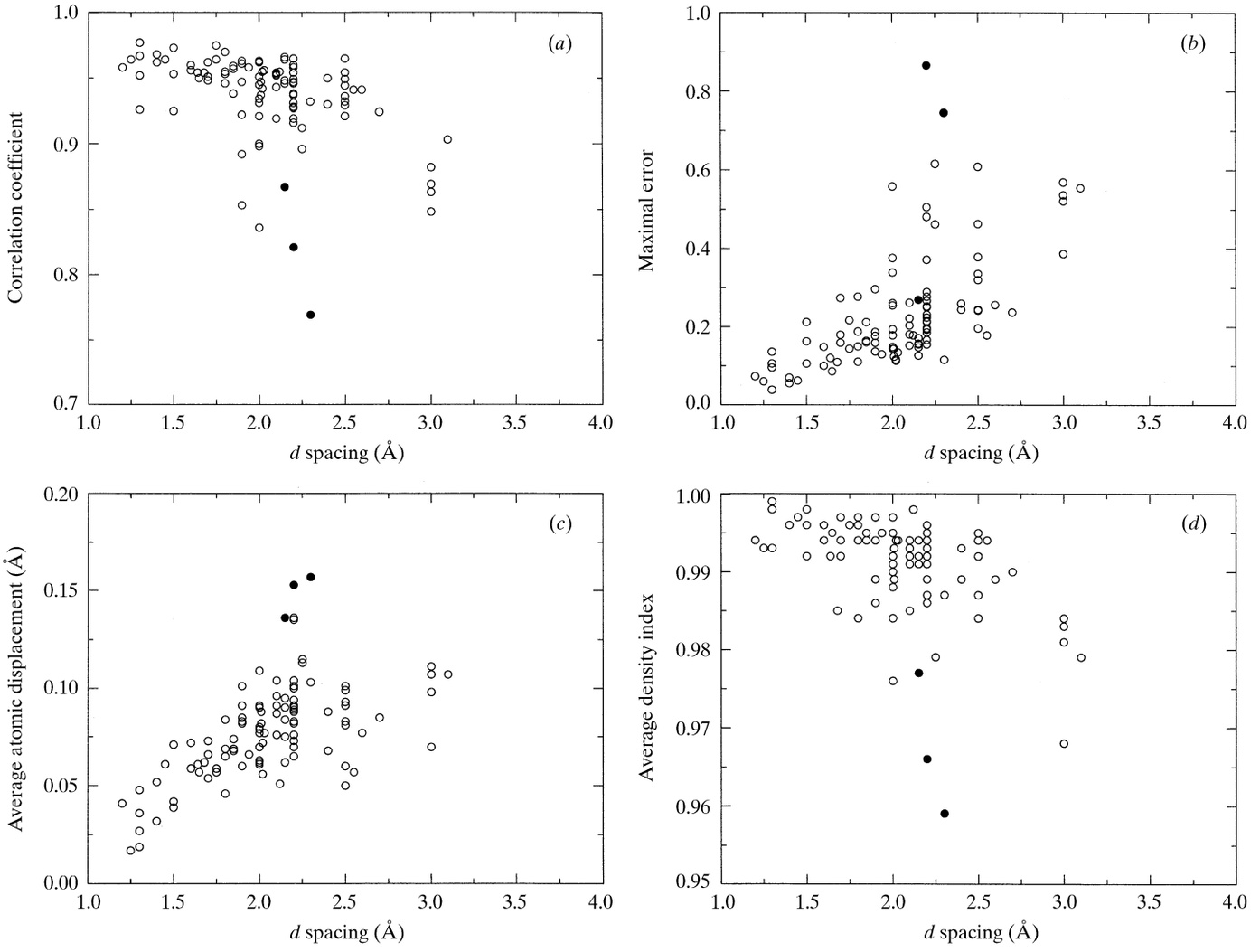


Fig. 21.2.3.4. Variation of global quality indicators with the nominal resolution (*d* spacing) of the crystallographic data. The following quality indicators were computed by *SFCHECK* for each of the 104 nucleic acid crystal structures considered in the study of Das *et al.* (2001): (*a*) correlation coefficient, (*b*) maximal error, (*c*) average atomic displacement and (*d*) average density index. For the meaning of the various quantities see Table 21.2.3.2. The three structures for which the reported and re-computed *R* factors differ by more than 10% are highlighted as black circles. The NDB (PDB) codes for these structures are ADFB72 (256D), ADF073 (257D) and ADJ081 (320D).

scaling procedures and refinement practices. Such biases are, however, well known for the *B* factors of individual atoms or residues. This is illustrated in Fig. 21.2.3.6(*a*). This figure plots, side-by-side, the average residue *B* factors in 21 protein structures determined at different *d* spacings. It shows that for proteins determined at poorer resolution (*d* spacing above 2 Å), the *B* factors of different structures are systematically shifted relative to one another. Such systematic shifts are much smaller for structures determined at 2 Å resolution or better (Fig. 21.2.3.6*a*). This is not surprising, since in lower-resolution structures, $N_{\rm refl}/N_{\rm atoms}$ is often too low (<4) to yield meaningful values for the *B* factors.

Interestingly, the residue Density index, a very different parameter from the *B* factor, which measures the level of electron density at the atomic positions, does not display the systematic shifts observed for the *B* factors (Fig. 21.2.3.6*b*), despite the fact that the two measures are rather strongly correlated in individual structures. An indicator such as this one, and ultimately the atomic s.u.'s themselves, should be better suited for analysing and comparing the trends in the quality of specific regions of the model across different structures.

21.2.4. Atomic resolution structures

With improved techniques of crystallization and data collection using synchrotron radiation and cryogenic cooling, an increasing number of protein crystal structures are being determined at atomic resolution (1.2 Å or better). With atomic resolution data, refinement can be performed that requires much less strict compliance with prior knowledge of the expected geometry. Although some restraints must still be imposed, especially to deal with more flexible regions, and hence biases remain, it might be expected that these structures provide more precise information on the 'true' geometrical and stereochemical properties of proteins. Ultimately, one would want to re-derive these properties using only atomic resolution structures, but their number is at present too limited to provide sufficient data for a meaningful statistical analysis.

In the meantime, atomic resolution protein structures have been used to check geometric and conformational parameters that have been derived from other sources, including small-molecule crystals and the larger set of proteins determined at various levels of